

# ИЗМЕРВАНЕ НА МЕЖДУЕЗИКОВА СЕМАНТИЧНА БЛИЗОСТ ЧРЕЗ ТЪРСЕНЕ В GOOGLE

Светлин Иванов Наков, Софийски университет „Св. Климент Охридски”,  
email: nakov @ fmi.uni-sofia.bg

## Резюме

В настоящата статия е описан алгоритъм за автоматично измерване на семантична близост между двойки думи на български и руски език. Алгоритъмът извлича локалните контексти на дадените думи чрез серия справки в търсещата машина Google и определя близостта между думите чрез сравнение на контекстите им. Извлечените локални контексти представляват честотни вектори, които носят семантична информация за най-честите думи, които се срещат в близост до дадена дума. Те могат да бъдат извлечени чрез серия заявки за търсене в Google и анализиране на върнатите като резултат отрязъци от текст. За преминаване от един език на друг се използва речник от двойки думи, превод една на друга.

## Ключови думи

Семантична близост, измерване на семантична близост, локален контекст, използване на уеб като корпус, Google.

## 1. Цели на изследването

Понятието „семантична близост” е мярка за това доколко две думи са свързани смислово. Близостта може да е изразена по най-различни начини: като синонимия (напр. хубав и прекрасен), като хипонимия (напр. дърво и липа) или като индиректна асоциация (напр. работилница и трион; бира и водка; космос и ракета; гозба и фурна).

Целите на настоящото изследване са да се разработи компютърен алгоритъм за автоматично измерване на семантична близост между двойки думи на различни езици (български и руски). Резултатите могат да бъдат използвани за решаване на различни задачи при обработката на естествен език: машинен превод, автоматично построяване на речници, различаване между няколко значения на дадена дума, моделиране на езика и други.

## 2. Алгоритъм за извличане на семантична близост чрез търсене в Google

Алгоритъмът за определяне на семантична близост от уеб е подобен на описания в [Наков и колектив, 2007b] и използва богатството на световната мрежа като източник на огромен брой текстове, от които се извлича семантична информация (използва се уеб като корпус).

Алгоритъмът изпълнява заявки в търсещата машина Google и анализира върнатите отрязъци от текстове. От тях извлича т. нар. *локален контекст* на всяка анализирана дума (думите в непосредствена близост до нея), тъй като той съдържа думи, които са семантично свързани с нея [Hearst, 1991]. По извлечените локални контексти за всяка дума се построява *честотен вектор*, който съдържа всички думи от съответните локални контексти заедно с честотите им на срещане. Семантичната близост между двойка думи се определя като косинус между честотните им вектори и представлява число между 0 и 1. Когато разглежданите думи са на различни езици, техните контексти (които също са на различни езици) се сравняват като предварително единият

контекст се преведе на другия език чрез речник. Алгоритъмът може да се ползва за измерване на семантична близост не само между думи, но и между фрази.

## 2.1. Семантична близост, измерена чрез контексти

Алгоритъмът за измерване на семантична близост между двойка думи се базира на концепцията за анализиране на локалния контекст, в който се срещат тези думи. Изхожда се от идеята, че думи, които се срещат в близък контекст, би трябвало да са близки по значение. Например думите *художник* и *картина* са семантично близки, тъй като се срещат най-често в изречения, в които става дума за художници, рисуване, картини, четки, бои, галерии и други термини, свързани с изобразителното изкуство.

Тъй като едно изречение може да е дълго, възниква въпросът каква част от него съдържа контекста на дадена дума. Обикновено лингвистите разглеждат т. нар. *локален контекст* на думата, а именно няколкото думи, които се срещат преди и след дадената дума. Да разгледаме за пример думата *картина* в следното изречение:

*Свидетелство за холандското изкуство по онова време е световноизвестната картина "Нощна стража", творение на майстора на играта със светлината Рембранд, който я рисува цели две години в периода 1640-1642 г.*

Локалният контекст на думата *картина* в това изречение съдържа думите, които се намират около нея (например 6 думи преди и 6 думи след нея), а именно: *изкуство, по, онова, време, е, световноизвестната, нощна, стража, творение, на, майстора, на*. Ако разгледаме основната форма на тези думи и премахнем функционалните думи, които не носят никаква семантика (предлози, местоимения, съюзи и т.н.), ще получим няколко думи, които представляват локалния контекст на думата *картина* в това изречение: *изкуство, време, световноизвестен, нощен, стража, творение, майстор*.

Някои от тези думи са семантично свързани с думата *картина*, но някои не са. Ако разгледаме достатъчно голям брой изречения с думата *картина* (примерно 1000) и извлечем от тях всички думи от нейния контекст, то най-често срещаните сред тях би трябвало да са най-типичните думи, които са семантично свързани с *картина*. Можем да очакваме сред тях да са думи като *художник, рисувам, четка, галерия и изкуство*. Думи като *време, нощен и стража*, случайно попаднали в контекста, не би трябвало да се срещат често, ако разгледаме достатъчно много изречения.

Нека сега вземем две думи и извлечем често срещаните думи от локалния им контекст, извлечен от достатъчно голям брой изречения. Ако думите са близки по значение, можем да очакваме техните контекстни думи и броят им срещания да съвпадат в голяма степен.

Описаното можем да формализираме като съпоставим вектори на срещанията (честотни вектори) за думите от контекстите на двете думи, които изследваме, и измерим косинуса на ъгъла между тези вектори. Например за думите *художник* и *картина* можем да имаме следните честоти на думите от техните локални контексти (в таблицата са със съкращения):

художник		картина	
художник	422	картина	461
картина	262	купувам	386
рисувам	202	скъп	345
изкуство	167	известен	205

галерия	94	галерия	183
известен	84	голям	176
купувам	72	изкуство	188
голям	56	художник	98
скъп	3	рисувам	91
фотоапарат	0	фотоапарат	2

**Таблица 1** – думи от локалните контексти на *художник* и *картина* и съответните им честоти

За размерности на векторите вземаме всички думи, които се срещат в контекстите на поне едната от двете думи, а за координати – честотите им на срещане. За думите, които не се срещат в даден контекст, приемаме, че се срещат 0 пъти. Получаваме следните два честотни вектора (в таблицата са със съкращения):

дума	вектор 1 (художник)	вектор 2 (картина)
художник	422	98
картина	262	461
рисувам	202	91
изкуство	167	188
галерия	94	183
известен	84	205
купувам	72	386
голям	56	176
скъп	3	345
фотоапарат	0	2

**Таблица 2** – контекстни честотни вектори за думите *художник* и *картина*

Доколко тези вектори си приличат можем да пресметнем като изчислим косинус в  $n$ -мерното евклидово пространство. Така получаваме число между 0 и 1, което дава числена стойност на семантичната близост между двете думи (по-голямо число означава семантично по-близки думи).

## 2.2. Семантична близост, измерена чрез уеб контексти

Световната мрежа съдържа най-големия в света корпус от текстове на различни езици, включително на български и руски език, което ни мотивира да я използваме като източник на информация с цел измерване на семантична близост между двойка думи. Ще опишем начин за извличане на локален контекст от уеб (*уеб контекст*), сходен с публикувания в [Nakov и колектив, 2007a].

За извличането на локалния контекст на дадена дума от Интернет използваме заявка за търсене на думата в уеб търсещата машина Google, в която указваме да бъдат върнати 100 резултата на съответния език (в нашия случай български или руски). С 10 такива заявки може да се извлекат до 1000 резултата (Google не позволява да извлечем повече). Всеки резултат съдържа заглавие и отрязък от текст, съдържащи търсената дума или нейна словоформа. Например за думата *картина* получаваме следния списък от заглавия и отрязъци от текст:

Нощна стража (**картина**) — Уикипедия

В момента <b>картината</b> е изложена в музея Рейксмузеум в Амстердам. Истинското име на <b>картината</b> е "Ротата на капитан Банинг Кок". Тъй като престояла дълги ...
<b>Картина</b> с известни личности   спанак.орг
Огромна <b>картина</b> , на която са изобразени много известни личности - Айнщайн, Чърчил, Линкълн, Фидел Кастро, Че Гевара. От новата вълна можете да намерите ...
Намерена е най-древната картина в света - MystiColors Forum
В будисткия комплекс Бамиян (Bamiyan) в Афганистан група японски археолози намериха най-древната в света <b>картина</b> , нарисувана с маслени бои. ...
...

Таблица 3 – резултати от търсене на думата *картина* в Google

Първо заменяме всички главни букви със съответните им малки и извличаме всички последователности от думи, като за разделител считаме всички знаци, които не са букви от кирилицата.

Следва премахване на всички функционални думи (предлози, местоимения, съюзи, частици, междуметия и някои наречия), както и думи с по-малко от 3 букви. Те не носят семантика и трябва да се пропуснат, защото изкривяват резултатите.

След това преминаваме през извлечените последователности от думи и търсим дадената дума (в нашия случай *картина*) или нейна словоформа и взимаме 3 думи преди и след нея (числото 3 наричаме размер на контекста). Тези думи считаме за част от локалния уеб контекст.

Всички извлечени думи заменяме с тяхната основна словоформа (прилагаме лематизация), например заменяме *картиной* с *картина*. За целта ползваме богат речник с лемите (за български и руски език).

Сега имаме всички думи, които се срещат в локалния уеб контекст на дадена дума и техните честоти на срещане (честотни вектори). Семантичната близост между двойка думи (на един и същ език) измерваме като пресметнем косинус между честотните им вектори в  $n$ -мерното евклидово пространство. Така получаваме число между 0 и 1, което показва доколко две думи си приличат семантично.

### 2.3. Междуетикова семантична близост

Измерването на семантична близост чрез контексти може да се приложи и за измерване на междуетикова семантична близост. За целта е необходим начин за сравняване на контексти на различни езици.

Нашият подход е базиран на идеята да се преведе първият контекст от единия език на другия и след това да се сравни с втория контекст по описания вече начин с измерване на косинуса на ъгъла между съответните честотни вектори.

За превеждане на думите от контекста от единия език на другия се използва речник. Речникът представлява съвкупност от двойки думи, превод една на друга. Тъй като от локалния контекст се извличат само единични думи, речникът не съдържа фрази. Когато за една дума от единия език има няколко съответни преводни думи от другия език, всяка от тях се взема под внимание с еднаква тежест. Думите от двата езика, за които няма съответно значение в речника, не се взимат под внимание.

Както руският, така и българският вектор винаги са с размер броя български думи в речника. Българският вектор пази за всяка българска дума в речника колко пъти се среща тя в контекста на българската дума. Руският вектор пази за всяка българска дума

от речника колко пъти се среща някой от нейните преводи в контекста на руската дума. Така се получават два вектора с еднаква размерност и близостта между тях може лесно да се измери.

## 2.4. TF.IDF претегляне

При извличане на информация (information retrieval) често пъти се прилага т. нар. TF.IDF претегляне на честотите на отделните думи. Числото TF.IDF (term frequency – inverse document frequency) е статистическа мярка, която измерва колко е важна дадена дума за даден документ от даден корпус с документи. Важността на думата се увеличава пропорционално спрямо броя на срещанията ѝ в документа, но намалява пропорционално на броя документи, които я съдържат. Установено е, че ако по-важните думи се взимат с по-голяма тежест, в крайна сметка се подобрява точността на резултатите от търсенето [Sparck-Jones, 1972].

За да приложим TF.IDF претегляне в алгоритъма за пресмятане на семантична близост правим следното: като получим първите 1000 резултата от търсене в уеб за дадена дума  $w$ , непосредствено изчисляваме честотите  $TF[w_i]$  на всички думи  $w_i$  в нейния контекст. Те се пресмятат като се раздели броя срещания на думата  $w_i$  на общия брой думи в контекста на  $w$  (с повторенията). След това изчисляваме  $IDF[w_i]$  като разделим общия брой документи, индексирани в Google (приемаме, че са 8 милиарда) на броя срещания в Google на думата  $w_i$ . Накрая взимаме  $\log_2(IDF[w_i])$  и умножаваме по  $TF[w_i]$  и така изчисляваме претеглената честота на думата  $w_i$  в честотния вектор на  $w$ . Полученият претеглен честотен вектор използваме за по-точно измерване на семантичната близост.

## 2.5. Семантична близост, измерена чрез обратен контекст

При извличане на локален контекст за дадена дума от уеб често пъти в него попадат думи, които не са семантично свързани с нея. Някои Интернет термини като *сайт*, *страница*, *блог*, *онлайн*, *форум*, *начало*, *връзки*, *меню*, *съобщение*, *изтегли* и др. се срещат в контекстите на почти всяка дума, но не са свързани семантично с нея. Премахването на такива думи от локалния уеб контекст следва да доведе до повишаване на точността при оценяване на семантичната близост, защото в контекста ще попадат само думи, които наистина имат семантична връзка с търсената дума [Nakov и колектив, 2007b].

Използването на обратния контекст се основава на идеята, че ако две думи са семантично свързани, то първата трябва да се среща често в контекста на втората и същевременно втората трябва да се среща често в контекста на първата.

Например в контекста на думата *картина* често се срещат думи като *художник*, *галерия* и *изкуство*, но и паразитни думи като *поръчвам*, *новини* и *сайт*. Ако направим търсене за първите три думи, ще се убедим, че *картина* се среща често в техните уеб контексти. Ако, обаче, направим търсене за последните три думи, ще се убедим, че в техните контексти *картина* почти не се среща.

Описаната идея може лесно да се формализира. Да означим с  $F(X, Y)$  броя срещания на думата  $Y$  в уеб контекста на думата  $X$ . Нека имаме думата  $A$  и извлечем думите от нейния уеб контекст  $W_i$  заедно с броя им срещания  $F(A, W_i)$ . Нека за всяка дума  $W_i$  извлечем броя срещания  $F(W_i, A)$  на думата  $A$  в нейния уеб контекст (обратен контекст). Получаваме вектор на взаимните срещания на думата  $A$  с всички думи от нейния контекст. Той е съставен от думите  $W_i$  с честоти:

$$\min(F(A, W_i), F(W_i, A))$$

Полученият вектор съдържа минималния брой взаимни срещания на дадена дума с всяка друга дума от нейния уеб контекст и съдържа по-точна семантична информация, отколкото чистия уеб контекст.

При изчисляването на вектора на взаимните срещания е добре да се игнорират думи, които се срещат прекалено малко на брой пъти (примерно по-малко от 10), защото това може да е случайно. С промяна на този параметър (*праг на честотата*) може да се влияе върху точността на резултатите.

## 2.6. Семантична близост чрез обогатяване на контекста

Често пъти в локалния контекст на дадена дума не попадат всички думи, които са семантично свързани с нея и я характеризират смислово. Например в контекста на думата *художник* е нормално да попадне думата *изкуство*, но различните видове изкуство, с които *художник* е свързана смислово, може и да не попаднат в нейния контекст. Логично е, ако думата *четка* е силно свързана смислово с *художник*, нейните синоними да се добавят в контекста на *художник* и така да го обогатят смислово.

Обогатяване на контекста означава да добавим към контекста на дадена дума контекстите на всички често срещани думи от нейния контекст [Nagiwara и колектив, 2007]. По този начин контекстът на думата се разширява с още думи, които оригинално не присъстват в него, но са свързани смислово с нея. Очакванията са обогатяванията на контекста да подобрят резултатите при измерване на междуетикова семантична близост, защото при повече думи работим с по-богата семантична информация.

При обогатяване на контекста, е добре да се игнорират думи, които се срещат в него прекалено малко на брой пъти (примерно по-малко от 10), защото това може да е случайно. С промяна на този параметър (*праг на честотата*) може да се влияе върху точността на резултатите, като се зададе разумна граница на минималния брой срещания, при който се извършва обогатяване на контекста.

## 3. Експерименти и резултати

Експериментите, които направихме, имат за цел да оценят предложените алгоритми за автоматично извличане на междуетикова семантична близост от уеб чрез сравнение на получените от тях резултати с оценки, дадени от човек.

### 3.1. Тестови данни

Като тестови данни използваме адаптация на списъка от 30 двойки думи, предложени от Милер и Чарлз [Miller & Charles, 1991]. Те представляват внимателно подбрани двойки съществителни имена, за всяка от които е направена оценка на семантичната близост от 51 души в скала от 0 до 4, след което оценката е усреднена.

Посочените от Милер и Чарлз 30 двойки думи преведохме съответно на български и руски език, като при многозначност се опитахме да запазим автентичността на оригиналните думи, доколкото е възможно. При превода на някои от думите използвахме фрази, за да предадем по-точно оригиналния смисъл. Преведените 30 двойки думи (и фрази) и съответните им човешки оценки за семантична близост са дадени в таблица 4:

#	Българска дума	Руска дума	Семантична близост оценена от човек (по Милер и Чарлз)
1	автомобил	автомобиль	3,92

2	скъпоценен камък	драгоценность	3,84
3	пътешествие	путешествие	3,84
4	момче	мальчик	3,76
5	крайбрежие	побережье	3,70
6	психиатрия	сумасшедший дом	3,61
7	магьосник	волшебник	3,50
8	пладне	полдень	3,42
9	Пещ	печь	3,11
10	Храна	фрукт	3,08
11	Птица	петух	3,05
12	Птица	журавль	2,97
13	инструмент	орудие	2,95
14	брат	монах	2,82
15	момче	брат	1,66
16	жерав	орудие	1,68
17	пътуване	автомобиль	1,16
18	калугер	оракул	1,10
19	гrobiща	лесистая местность	0,95
20	храна	петух	0,89
21	крайбрежие	холм	0,87
22	гора	кладбище	0,84
23	бряг	лесистая местность	0,63
24	калугер	раб	0,55
25	крайбрежие	лес	0,42
26	момче	волшебник	0,42
27	корда	улыбка	0,13
28	стъкло	маг	0,11
29	петел	путешествие	0,08
30	пладне	нитка	0,08

Таблица 4 – адаптация за български и руски език на 30-те двойки думи на Милер и Чарлз

При превода не винаги намерихме точно съответствие между английски, български и руски език и на много места са изгубени нюансите на оригиналните думи, което съответно прави неточна дадената човешка оценка на семантичната близост.

### 3.2. Използвани ресурси

За целите на експериментите и при реализирането на алгоритъма за извличане на семантична близост от уеб бяха използвани следните ресурси:

- **Online уеб търсеща машина Google<sup>1</sup>**. Използвана е информацията за първите 1000 резултата при търсене на 30 943 български и 32 645 руски думи.
- **Граматичен речник на българския и руския език**, изготвен в Лабораторията за лингвистично моделиране към Института за паралелна обработка на информация

<sup>1</sup> <http://www.google.com>

към Българска академия на науките (БАН) [Paskaleva, 2007]. В българския си вариант речникът съдържа 963 339 словоформи и 73 113 леми. В руския си вариант речникът съдържа 1 390 613 словоформи и 66 101 леми. Всяка речникова единица съдържа словоформа, съответната ѝ лема, следвани от морфологична и граматична информация.

- **Списък с функционалните думи в българския и руския език.** Съдържа съответно 598 български и 507 руски думи (предлози, местоимения, съюзи, частици, междуметия и наречия). Съставен ръчно.
- **Кратък българо-руски речник,** съдържа 4 562 двойки думи, които са превод една на друга. Съставен от онлайн българо-руски речник [BgRu.net, 2007].
- **Подробен българо-руски речник,** съдържа 59 582 двойки думи и фрази, които са превод една на друга. Съставен от два големи българо-руски и руско-български речника [Чукалов, 1986] и [Бернщайн, 1986].

### 3.3. Описание на експериментите

Върху 30-те думи и фрази от таблица 4 са проведени серия експерименти за оценяване на семантичната им близост чрез изпълнение на описаните алгоритми при различни техни параметри (с различна големина на речниците, с и без прилагане на TF.IDF, с и без използване на обратен контекст, с и без обогатяване на контекста и при различни стойности на минималната честота на срещане на думите).

Получените резултати от нашите алгоритми за автоматичното измерване на семантична близост са сравнени с човешката оценка чрез изчисление на *коефициент на корелация на Пирсън*. Този коефициент е стандартна статистическа мярка за измерване на линейна взаимовръзка между две променливи величини, която показва доколко те са близки една с друга. Стойността на корелационния коефициент на Пирсън е между -1 и 1 като стойности близки до 0 означават липса на близост, а стойности близки до -1 или 1 означават висока степен на близост.

В нашия случай високата корелация (стойности близо до 1) означава голяма близост между машинно извлечените и човешките оценки за семантична близост, т.е. един алгоритъм е толкова по-добър, колкото по-висока е корелацията му с човешките оценки за нашите 30 двойки думи.

Върху тестовите данни са извършени следните експерименти:

- **RAND** – случайна близост, зададена за всички двойки думи. Ползва се като база за сравнение с другите алгоритми.
- **SIM** – основният алгоритъм за извличане на семантична близост от уеб (описан подробно в [т. 2.1](#), [2.2](#) и [2.3](#)), с краткия българо-руски речник, с размер на контекста 3 думи, без използване на обратен или обогатен контекст, с прилагане на лематизация.
- **SIM-BIG** – основният алгоритъм SIM с подробния българо-руски речник.
- **SIM+TFIDF** – модификация на SIM алгоритъма с използване на TF.IDF претегляне (описан подробно в [т. 2.4](#)).
- **SIM-BIG+TFIDF** – модификация на SIM алгоритъма с използване на TF.IDF претегляне и използване на подробния българо-руски речник.



- **REV-0, REV-10, REV-20, REV-30, REV-40, REV-50** – модификация на SIM алгоритъма с използване на обратен контекст (описан подробно в [т. 2.5](#)) с прагове на честота съответно 0, 10, 20, 30, 40 и 50.
- **REV-BIG-0, REV-BIG-10, REV-BIG-20, REV-BIG-30, REV-BIG-40, REV-BIG-50** – модификация на SIM алгоритъма с използване на обратен контекст с прагове на честота съответно 0, 10, 20, 30, 40 и 50 и с подробния българо-руски речник.
- **IND-10, IND-20, IND-30, IND-40, IND-50** – модификация на SIM алгоритъма с използване на обогатен контекст (описан подробно в [т. 2.6](#)) с прагове на честота съответно 10, 20, 30, 40 и 50.
- **IND-BIG-10, IND-BIG-20, IND-BIG-30, IND-BIG-40, IND-BIG-50** – модификация на SIM алгоритъма с използване на обогатен контекст с прагове на честота съответно 10, 20, 30, 40 и 50 и с подробния българо-руски речник.

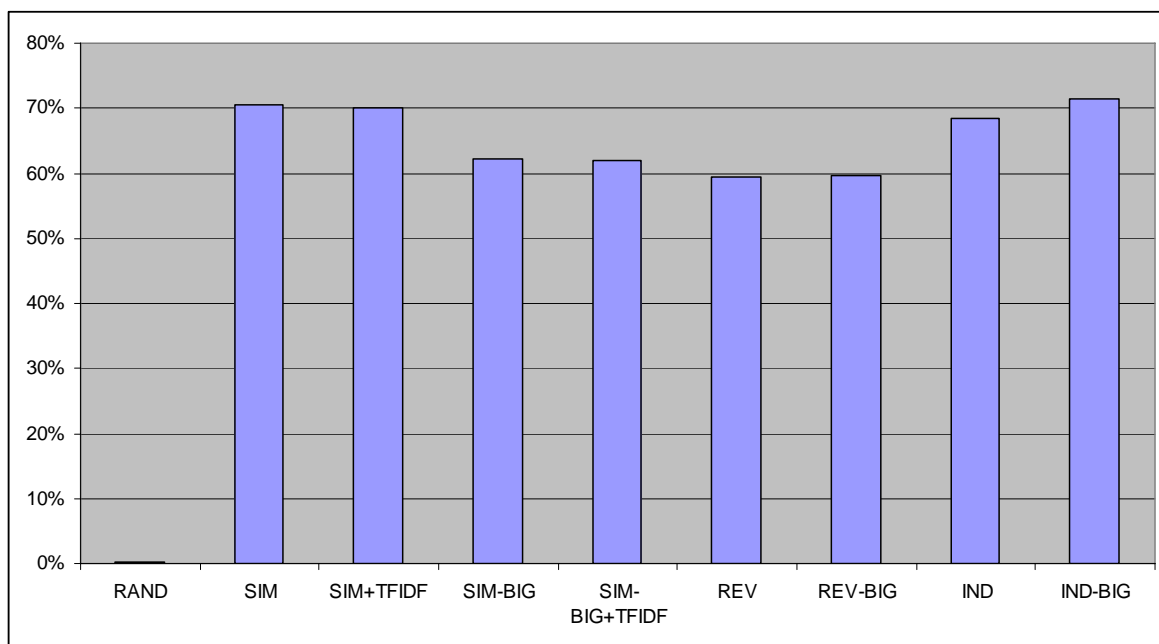
### 3.4. Резултати

Таблица 5 показва резултатите от проведените експерименти:

Алгоритъм	Праг 0	Праг 10	Праг 20	Праг 30	Праг 40	Праг 50
RAND	<b>0,0000</b>	-	-	-	-	-
SIM	<b>0,7043</b>	-	-	-	-	-
SIM+TFIDF	<b>0,7010</b>	-	-	-	-	-
SIM-BIG	<b>0,6210</b>	-	-	-	-	-
SIM-BIG+TFIDF	<b>0,6191</b>	-	-	-	-	-
REV	<b>0,5933</b>	0,5732	0,5623	0,5625	0,5623	0,5492
REV-BIG	0,5961	<b>0,5964</b>	0,5956	0,5957	0,5953	0,5920
IND	-	0,5078	0,6027	<b>0,6850</b>	0,6485	0,6445
IND-BIG	-	0,5046	0,6057	<b>0,7149</b>	0,6296	0,6412

Таблица 5 – резултати от различните алгоритми (корелация на Пирсън)

В нея са дадени получените стойности на корелационния коефициент на Пирсън между семантичната близост, измерена от различните алгоритми, и семантичната близост оценена от човек. В получер шрифт са дадени най-високите стойности за различните прагове на съответния алгоритъм, които са изобразени и на диаграмата:



Диаграма 1 – сравнение на резултатите от различните алгоритми

### 3.5. Анализ на резултатите

От таблица 5 е видно, че семантичната близост, оценена автоматично с предложените алгоритми, има корелация със съответните човешки оценки от 50% до 71%. Тази корелация е много по-висока от 0%, която се получава при случайната оценка RAND.

Макар и резултатите от основния алгоритъм SIM да са доста добри, виждаме, че всички опити за неговото подобрене не са много успешни. От резултатите можем да направим следните заключения:

- TF.IDF претеглянето влияе негативно.
- Използването на обратен контекст не работи добре и REV алгоритъма работи по-лошо от основния SIM алгоритъм при всякакви прагове на честотата.
- Обогащаването на контекста (IND алгоритъма) работи малко по-добре от основния SIM алгоритъм само при внимателно подбран праг на честотата.
- Използването на подробния вместо краткия речник помага само в някои случаи.

Точността на получените резултати при никой алгоритъм не е 100%. Основните причини за това са няколко:

- Неточност при превода на 30-те думи на Милер и Чарлз. При междуетиковия превод не винаги има точно съответствие между английски, български и руски език и на много места са изгубени нюансите на оригиналните думи. Това прави човешката оценка на семантичната близост между посочените думи неточна и е основна причина за намаляване на измерената корелацията.
- Съществува неточност при измерване на семантична близост чрез предложените алгоритми, защото не винаги думи, които се срещат в близък контекст са близки по значение. Някои думи попадат в контекста случайно или имат много значения, само едно от които е семантично свързано с търсената дума.
- Използването на уеб като корпус ограничава извличането на локален контекст измежду само 1000 статии, а те не са представителна извадка на всички статии.

Често пъти новинарски и търговски сайтове се позиционират в челни позиции в резултатите при търсене в Google за сметка на други текстове.

- Използването на думи, вместо фрази в контекстите и след това при превода внася много шум. Много от думите имат по няколко значения, а това означава неточен превод и съответно неточни резултати при сравнение на контекстите.
- Непълнота на преводните речници. Думите от двата езика, за които няма съответно значение в речника, не се взимат под внимание дори ако се срещат много голям брой пъти. Това предизвиква неточности при някои по-рядко срещани думи.

#### **4. Други разработки по темата и сравнение с тях**

Повечето известни методи за автоматично оценяване на семантична близост се базират в някаква степен на лингвистичната хипотеза за разпределението (distributional hypothesis) [Harris, 1954], която твърди, че семантично близките думи се срещат в близки контексти. Нашият метод също се основава на тази хипотеза (извличане на контексти и тяхното сравнение), но ние използваме уеб като корпус за извличане на контекстите и така методът ни не зависи от наличието на други лингвистични ресурси (например големи корпуси за дадения език).

Алгоритми, базирани на хипотезата за разпределението са предложени в [Lin, 1998] и [Curran & Moens, 2002]. При тях контекстите се дефинират на базата на предефинирани граматични релации, които се извличат от едноезичен корпус. В крайна сметка отново се измерва близостта между извлечените контексти.

Идеята да се използва уеб като корпус е използвана от много изследователи при решаването на различни задачи [Kilgarriff & Grefenstette, 2003]. Някои използват уеб търсещи машини за намиране на броя срещания на дадена дума или фраза и изчисляване на взаимна информация (pointwise mutual information) [Inkpen, 2007], докато други директно извличат контекст от откъсите текст, които уеб търсещите машини връщат [Nakov и колектив, 2007b].

Идеята за извличане на информация от откъсите текст, връщани от уеб търсещите машини е използвана в [Chen и колектив, 2006]. Предложеният от тях модел WSDC се базира на идеята, че ако две думи X и Y са семантично свързани, то при търсене на X в резултатите ще се появява често думата Y и обратното - при търсене на Y в резултатите ще се появява често думата X. При този подход напълно се игнорират контекстните думи (освен X и Y) и не се използва семантиката, която те носят. Както по-късно е установено [Bollegala и колектив, 2007], това води до некоректна нулева семантична близост за огромна част от двойките думите, които се разглеждат.

[Bollegala и колектив, 2007] комбинират извличането на информация за броя срещания на дадени две думи (заедно и поотделно) от уеб търсеща машина с извличане на информация от откъсите текст, върнати от търсещата машина. При техния подход предварително се извличат автоматично лексично-граматични шаблони за семантично свързани и несвързани думи чрез използване на WordNet и се тренира support vector machine (SVM). След това заучените шаблони се използват за извличане на информация от откъсите текст, върнати от търсещата машина. Накрая резултатите се комбинират. Оценката се прави спрямо други методи като се използват едни и същи тестови данни на английски език. Методът е по-сложен от нашия и изисква допълнителни ресурси, които се използват при тренирането на SVM. Не сме направили сравнение с него, тъй като нашите тестови данни са на руски език.

Огромно предимство на предложения от нас метод е, че не изисква големи корпуси или други ресурси (като например WordNet), които за някои езици може да не са налични. Това дава основание да очакваме, че той може много успешно да се комбинира с други подходи и да им подобри точността, особено в случаите, когато има недостиг на ресурси.

[Budanitsky & Hirst, 2006] сравняват 5 различни алгоритъма за оценяване на семантична близост, базирани на WordNet. Най-добрият от тях постига корелация на Пирсън от 85% за 30-те двойки думи на Милер и Чарлз, което е по-добър резултат в сравнение с нашите алгоритми, но използва семантичната мрежа WordNet, която няма български и руски вариант.

[Weeds, 2003] сравнява 6 алгоритъма за извличане на семантична близост, базирани на хипотезата за разпределението (и неизползващи допълнителни ресурси), и установява, че най-добрият от тях постига коефициент на корелация на Пирсън от 62%. Нашият резултат е 71% и се отнася за по-сложната задача (измерване междуезикова семантична близост) е далеч по-добър от описаните от Weeds алгоритми. Това показва, че използването на уеб като корпус е сериозна стъпка напред при измерването на семантична близост без използването на допълнителни семантични ресурси (като WordNet).

## 5. Библиография

[Hearst, 1991] Hearst M. "Noun Homograph Disambiguation Using Local Context in Large Text Corpora". In *Proceedings of the 7th Annual Conference of the University of Waterloo Centre for the New OED and Text Research*, Oxford, England, 1991, pages 1-22.

[Nakov и колектив, 2007a] Nakov P., Nakov S., Paskaleva E. "Improved Word Alignments Using the Web as a Corpus". In *Proceedings of RANLP'2007*, pages 400-405, Borovetz, Bulgaria, 2007.

[Nakov и колектив, 2007b] Nakov S., Nakov P., Paskaleva E. "Cognate or False Friend? Ask the Web!". In *Proceedings of the Workshop on Acquisition and Management of Multilingual Lexicons, held in conjunction with RANLP'2007*, pages 55-62, Borovetz, Bulgaria, 2007.

[Sparck-Jones, 1972] Sparck-Jones K. "A Statistical Interpretation of Term Specificity and its Application in Retrieval". *Journal of Documentation*, Volume 28, 1972, pages 11-21.

[Miller & Charles, 1991] Miller, G., Charles W., Contextual Correlates of Semantic Similarity, *Language and Cognitive Processes*, 1991, 6(1):1-28

[Hagiwara и колектив, 2007] Hagiwara M., Ogawa Y., Toyama K. (2007). "Effectiveness of Indirect Dependency for Automatic Synonym Acquisition". In *Proceedings of CoSMo 2007 Workshop, held in conjunction with CONTEXT 2007*, Roskilde, Denmark.

[Paskaleva, 2002] Paskaleva E. "Processing Bulgarian and Russian Resources in Unified Format". In *Proceedings of the 8th International Scientific Symposium MAPRIAL*, Veliko Tarnovo, Bulgaria, 2002, pages 185-194.

[BgRu.net, 2007] Online Bulgarian-Russian and Russian Bulgarian dictionary – <http://www.bgru.net/intr/dictionary/>.

[Чукалов, 1986] Чукалов С. К. "Руско-български речник", Издателство "Русски език", Москва, 1986

[Бернщайн, 1986] Бернщайн, С. Б. "Българо-руски речник". Издателство "Русски език", Москва, 1986

[Harris, 1954] Harris, Z. "Distributional structure". *Word*, volume 10, 1954, pages 146-162.

- [Lin, 1998] Lin D. (1998). "Automatic Retrieval and Clustering of Similar Words". In *Proceedings of COLING-ACL'98*, Montreal, Canada, pages 768-774.
- [Curran & Moens, 2002] Curran J., Moens M. "Improvements in Automatic Thesaurus Extraction". In *Proceedings of the Workshop on Unsupervised Lexical Acquisition, SIGLEX 2002*, Philadelphia, USA, pages 59-67.
- [Kilgarrieff & Grefenstette, 2003] Kilgarrieff A., Grefenstette G. "Introduction to the Special Issue on the Web as Corpus", *Computational Linguistics*, 2003, 29(3):333–347.
- [Inkpen, 2007] Inkpen D. "Near-synonym Choice in an Intelligent Thesaurus". In *Proceedings of the NAACL-HLT*, New York, USA, 2007.
- [Chen и колектив, 2006] Chen H., Lin M., Wei Y. "Novel Association Measures Using Web Search with Double Checking". In *Proceedings of the COLING/ACL 2006*, Sydney, Australia, pages 1009-1016.
- [Bollegala и колектив, 2007] Bollegala D., Matsuo Y., Ishizuka M. "Measuring Semantic Similarity between Words Using Web Search Engines". In *Proceedings of the 16th International World Wide Web Conference (WWW2007)*, Banff, Canada, 2007, pages 757-766.
- [Budanitsky & Hirst, 2006] Budanitsky A., Hirst G. "Evaluating WordNet-based Measures of Lexical Semantic Relatedness". *Computational Linguistics*, Volume 32, Issue 1, 2006, MIT Press, USA, pages 13-47
- [Weeds, 2003] Weeds J. "Measures and Applications of Lexical Distributional Similarity", Ph.D. Thesis, University of Sussex, 2003